中国科学: 信息科学 2018 年 第 48 卷 第 11 期: 1521–1532
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
社会媒体计算与自然语言处理专刊 • 论文

融合知识表示的知识库问答系统

安波[1,2*], 韩先培[2], 孙乐[2]

1. 中国科学院大学, 北京 100190
2. 中国科学院软件研究所, 北京 100049
* 通信作者. E-mail: anbo@iscas.ac.cn

收稿日期: 2018–08–10; 接受日期: 2018–10–22; 网络出版日期: 2018–11–15

国家自然科学基金 (批准号: 61433015, 61572477, 61772505) 和中国科协青年人才托举工程 (批准号: YESS20160177) 资助项目

摘要 基于知识库的问答系统能够根据知识库中的事实自动回答自然语言的问题. 简单问题是指可以通过知识库中单一的事实来进行回答的问题, 这类问题也是最常见的问题. 但是当面对大规模的知识库时, 简单问题依然存在很大的挑战. 当前的端到端 (end-to-end) 模型主要依赖于对问句、实体和关系的文本描述进行表示学习, 进而根据这些表示来计算实体、关系与问句的语义相关度, 忽略了知识库中的实体和关系的结构信息. 而这些结构信息, 对于问句中实体和关系的识别有重要作用. 本文采用一种融合文本和知识的表示学习方法, 通过文本表示和组合模型来学习问句和知识的表示, 同时使用知识的结构信息来约束文本的表示和组合. 在基于知识的问答任务上的结果表明, 本文提出的方法学习到的问句和知识的表示能很好地反映问句与知识之间的语义相关性, 并显著地提升了问句中实体链接和关系识别的准确率.

关键词 问答系统, 知识库, 文本组合, 知识表示, 文本表示

1 引言

近年来, 知识库取得了快速的发展, 出现了许多大规模的知识库, 例如 Freebase [1], YAGO [2] 和 WordNet [3]. 这些大规模知识库, 也衍生出了很多自然语言处理 (NLP) 的研究方向和应用, 比如基于知识库的问答系统 (KBQA). 典型的知识库采用三元组来存储事实 $\{(h, r, t)\}$, 其中 h, r, t 分别表示头部实体、关系和尾部实体. 比如 ("Barack Obama", "people/person/spouse_s./people/marriage/spouse", "Michelle Obama").

基于知识库的问答系统能够自动地使用知识库中的事实信息来回答自然语言的问题. 其核心在于理解自然语言的问句, 并将问句中所涉及的实体和关系识别出来. 比如给定问题 "Who is Yao Ming's wife?", 想要得到该问题的答案则需要首先识别出实体 "Yao Ming"、关系 "wife". 但是由于知识库中

引用格式: 安波, 韩先培, 孙乐. 融合知识表示的知识库问答系统. 中国科学: 信息科学, 2018, 48: 1521–1532, doi: 10.1360/N112018-00208
An B, Han X P, Sun L. Knowledge-representation-enhanced question-answering system (in Chinese). Sci Sin Inform, 2018, 48: 1521–1532, doi: 10.1360/N112018-00208

只有 “spouse” 关系, 而没有 “wife” 关系, 因此还需要将 “wife” 对应到 “spouse” 这个关系. 当识别出问句中涉及的实体和关系之后, 则可以通过查询知识库中的三元组 (“Yao Ming”, “people/person/spouse_s./people/marriage/spouse”, “Ye Li”) 得到答案 “Ye Li”.

基于知识库的问答系统的实现方法主要分为 3 类: 基于语义解析的方法、基于信息检索的方法和基于深度学习的方法 [4~9]. 基于语义解析的方法依赖于人工定义的逻辑表达式; 基于检索的方法依靠复杂的特征工程; 基于深度学习的方法不需要人工定义逻辑表达式以及复杂的特征工程, 而是采用了端到端的方式来直接得到问句中包含的实体和关系, 因此成为了近期研究的热点. 目前的大多数基于知识库的问答系统在学习问句、实体和关系表示的时候主要是基于实体/关系的文本描述信息, 而忽略了知识库的结构信息. 如, 给定一个实体 “Yao Ming”, 常用的方法是通过实体名 (name)、实体类型 (type) 等描述信息来学习其表示. 这些方法忽略了知识库本身的结构信息, 而这些结构信息不仅可以学习到实体/关系的语义, 还可以对于实体和关系的组合进行有效的约束. 例如, 当问题中的关系为配偶关系时, 则其涉及的实体应该是人物. 而知识库中的结构信息, 因对实体和关系的组合进行了语义的约束, 可以避免识别出的实体和关系出现语义不匹配的情况.

近期, Hao 等 [10] 提出使用知识库中的结构信息来增强问答系统, 通过 TransE [11] 模型学习实体和关系的表示, 然后与基于实体的文本描述得到的表示进行拼接, 作为实体的表示. 该方法在知识的表示学习中将实体和关系作为一个独立的单元, 得到的表示完全基于知识库的结构信息. 但是学习到的实体和关系的表示与文本表示之间存在语义鸿沟, 不能直接提升实体/关系的表示.

针对该问题, 本文提出一种融合知识表示的知识库问答系统, 该模型利用知识库的结构信息来提高识别问句中的实体/关系的准确率. 具体地, 在知识库的表示学习过程中, 不再将实体/关系看作独立的单元, 而是通过对文本的表示和组合得到实体/关系的表示. 然后利用知识库中的三元组更新文本的表示和组合模型, 从而实现使用知识库的结构信息来增强文本的表示和组合模型. 本文使用基于卷机神经网络学习字符级别的表示和组合 (CharCNN) 得到词的表示. 组合得到的词表示与从词本身的向量表示相加, 作为词最终的表示. 词的最终表示在问句表示和实体/关系的表示中是通用的. 问句、实体和关系的表示基于双向长短时记忆网络 (Bidirection long short-term memory, BiLSTM) 实现, 每个词的最终表示作为双向长短时记忆网络的输入, 通过对词表示向量进行组合可以得到句子级别的表示. 由于一个问句可能涉及多个候选实体, 且问句中不同的单词对于实体/关系识别的重要程度是不同的. 因此本文引入注意力机制 (attention), 学习问句中每个单词对于不同实体的权重同时, 实体和关系的识别是相互影响的, 比如识别出是 “people/person/spouse_s./people/marriage/spouse” 关系, 则实体的类型应该是 “person”. 本文同时识别问句中的实体和关系, 并根据最终地识别结果来计算模型的损失. 联合训练的方式也有助于验证知识表示对于实体和关系识别的作用. 本文使用一种联合训练的方式来训练模型, 迭代地进行知识的表示学习和实体链接/关系识别.

本文在公开数据集 (SimpleQuestion [7]) 上来验证我们的模型. 实验表明, 本文提出的方法取得了最好的效果, 并且验证了知识表示、注意力机制、迭代训练等模块对模型地提升作用. 本文的主要贡献包括: (1) 提出一种融合文本和知识的表示方法, 该方法能将文本和知识学习到统一的表示空间; (2) 验证了字符组合表示可以提升词级别表示质量, 进而提高实体/关系识别的准确率; (3) 发现通过迭代地进行知识表示和问句的实体/关系识别的学习, 能够有效提升模型的准确率.

2 相关工作

本文利用知识表示来增强基于知识库的问答系统, 其中问句和实体/关系的表示是基于文本表示

及组合得到的. 因此本文主要涉及的相关工作包含 3 类: 文本的表示和组合模型、知识表示学习、基于知识库的问答系统.

文本表示及组合. 当前大多数的文本表示是基于神经网络或者共现矩阵 (co-occurrence) 学习得到的. 词级别的表示学习常用的模型包括 C&W [12], Word2Vec [13] 和 GloVe [14] 等. 当前也有一些模型使用字符级别 (character-level) 的表示进行组合得到词表示, 如基于 CNN 的字符组合模型 [15]、基于 BiLSTM 的字符组合模型 [16] 等. 基于字符组合的词表示模型能够避免集外词 (out-of-vocabulary, OOV) 问题, 并且可以利用词内部的构成信息, 是对词向量的有效补充.

对于短语句子的表示学习, 不能采用词表示的学习方法, 因为会出现严重的数据稀疏问题. 因此语义组合成为了一种可行的长文本表示方法. 语义组合是当前的研究热点, 其主要方法包括简单的词向量相加/点乘的方法 [13,17]、基于神经网络的方法 (如基于循环神经网络 (recurrent neural network) 的方法 [18]、基于卷机神经网络的方法 [8]、基于长短时记忆网络 (LSTM) [19] 的方法和基于注意力机制的方法 [20]).

本文提出的模型利用 BiLSTM 得到句子/实体/关系的表示, 同时使用了字符级的组合模型和词级的组合模型. 能够比较好地处理实体名中的单词出现次数较少的情况.

知识表示学习. 知识库表示学习的目标主要是根据知识库中三元组的结构信息学习到实体和关系的表示, 让实体和关系的表示去拟合知识的结构语义. 知识表示模型可以分为完全基于结构的模型和文本及路径增强的模型. 基于结构的模型主要是通过损失函数来更新实体或关系的表示, 如 unstructured model [21] 是让头部实体 h 和尾部实体 t 尽可能的相近. Distance model [22] 是在 unstructured model 的基础上增加两个关系映射矩阵, 分别将头部实体和尾部实体映射到关系向量空间中, 加入了关系的表示. Single layer model [23] 模型在 distance model 的基础上增加了单层神经网络, 增强了模型的表示能力. Bilinear model [24] 模型使用一个单独的矩阵来表示关系, 一个向量来表示实体, 其损失函数转换为三元组中实体和关系矩阵的相乘. DistMult [25] 模型使用实体和关系向量的点乘建模三元组. ComplEx [26] 模型是对 DistMult 模型的扩展, 将向量从实数空间扩展到虚数空间, 这样能够有效地避免三元组中一对多、多对多等关系的表示能力. Neural tensor network [23] 模型是对 single layer model 和 bilinear model 模型的结合, 增加了模型的表示能力, 但是同时也增加了模型的参数, 因而对训练数据的需求较高, 需要的计算时间也较长. TransE [11] 是近期提出的一种简单高效的模型, 该模型将三元组看做是在向量空间中从头部实体通过关系到尾部实体的转移, 后续有很多模型都是针对该模型的改进. 比如, TransH [27], TransM [28], TransR [29] 等是将实体表示和关系表示放在不同的语义空间, 通过矩阵进行映射. 同时也有一些模型通过文本、逻辑规则和路径等信息来增强知识的表示学习. PTransE [30] 使用路径来增强知识库的表示学习. 此外逻辑规则也被用于增强知识表示 [31~34].

本文的知识表示学习是基于 TransE 模型实现的, 但是我们的方法也可以使用其他的知识库表示学习模型, 因此可以随着知识库表示学习模型的发展, 得到进一步的提升.

基于知识库的问答系统. 基于知识的问答系统主要分为 3 类方法, 基于语义解析的方法、基于检索的方法和基于深度学习的方法. 基于语义解析的方法首先将自然语言的问句转换为适合在知识库中检索形式的逻辑表达式, 比如 CCG [35] 文法、Lambda 表达式 [36] 等. 然后使用得到的逻辑表达式到知识库中查询得到最终的答案. 基于信息检索的方法首先通过字符匹配等方法得到一系列的候选实体和关系, 然后通过特征对得到的结果进行排序, 排序最高的即为最终的结果. 基于深度学习的方法成为近期的主流方法, 该类方法基于神经网络将问句、实体和关系转换为向量表示, 进而在向量空间中识别实体和关系. Border 等 [7] 提出的基于记忆网络的方法学习词袋的表示, 并且验证了外部资源对于基于知识库的问答系统作用. Golub 等 [6] 提出使用字符级别的表示和基于注意力机制的自编码模型来学

习句子的表示. Yin 等 [8] 提出使用卷积神经网络 (CNN) 和注意力机制来改进模型, 该模型使用 CNN 来作为词级别的组合模型. Lukovnikov 等 [37] 近期提出了使用门循环神经网络 (gated recurrent unit, GRU) 来组合字符级别的表示得到词表示, 然后使用 GRU 来组合词表示得到句子级别的表示. 上述的各种方法均是利用文本信息来学习句子和知识库中实体和关系的表示. 尽管这些方法已经取得了比较好的效果, 但却忽略了知识库本身提供的结构语义信息. 而知识库中的结构语义信息包含了实体之间的相互关系, 以及实体与关系的隐含约束, 如 "parentOf" 关系对应的实体类型为 "person". 这些结构语义信息对于实体和关系的识别和问句的表示都有重要价值. Hao 等 [10] 提出使用知识库的结构信息来增强基于知识的问答系统, 该方法将实体/关系看作独立的单元, 学习其表示. 然后将这个表示与通过文本组合得到的表示进行拼接, 作为实体和关系的表示. 但是该方法不能直接用知识库的结构信息来增强文本的表示和组合模型, 忽略了文本和知识存在的语义鸿沟 (semantic gap).

本文提出一种融合知识表示的基于知识库的问答系统, 能够利用知识库的结构信息来增强文本的表示和组合模型. 本方法在进行知识库的表示学习的时候不是将实体和关系看做一个独立的单元, 而是通过文本的表示和组合方法来得到实体/关系的表示. 本文提出的方法使用 CharCNN 来进行字符级的组合来学习词的表示, 然后基于 BiLSTM 学习知识库中实体/关系的表示. 学习到的表示用于知识表示的训练, 对文本的表示和组合模型进行更新, 进而将知识库的结构信息编码到文本表示和组合模型中. 最终得到的实体/关系表示用于监督问句的表示学习, 从而得到更符合知识库语义和约束的表示.

3 融合知识表示的知识库问答系统

本文主要解决的是简单类型问题, 该类型的问答系统主要包括实体链接、关系识别、答案检索 3 个主要步骤: (1) 实体链接. 识别出问句中所涉及的主要实体. 形式化定义为: 给定一个包含 n 个单词的问句 $Q = w_1, w_2, \ldots, w_n$, 找出其中表示实体的字符串并链接到知识库中的实体 e_i; (2) 关系识别. 判断问句的关系词, 并将关系词匹配到知识库中的关系 r_j; (3) 答案检索. 根据找到的实体和关系到知识库中查询, 找到目标结果, 也就是给定 e_i 和 r_j 到知识库中找到匹配的三元组. 当完成问句中实体和关系的识别之后, 答案的检索只需要到知识库中查询即可. 因此本文主要解决问句的实体链接和关系识别, 不涉及最后答案的检索.

本文提出的方法主要包括 4 个模块: 候选实体和关系的生成、融合知识表示的文本表示及组合模型、实体链接, 以及关系识别.

3.1 融合知识表示的文本表示及组合模型

本小节主要介绍如何将知识表示所包含的结构信息编码到文本表示和组合模型中. 针对知识库中的每一个实体抽取其实体名 (name) 和其类别的文本名 (type), 比如: Freebase 中的实体 "m.0j3yffd" 的文本名和类别名分别为 "Terry Adamson" 和 "Football player..". 关系则使用关系本身的文本描述信息, 比如 "people/person/nationality" 转换为词序列 "people person nationality". 这样我们将知识库中的三元组转换为文本形式的三元组, 如: (m.0j3yffd, people/person/nationality, m.0162v) 转换为 "(Terry Adamson (Football player..), people person nationality, Barbado (Country))".

对于每个给定的文本三元组, 首先使用文本表示和组合模型得到每个实体和关系的表示. 实体/关系的表示基于字符级别的组合表示得到, 其中词表示如图 1 所示. 该模型通过充分利用词的内部组成

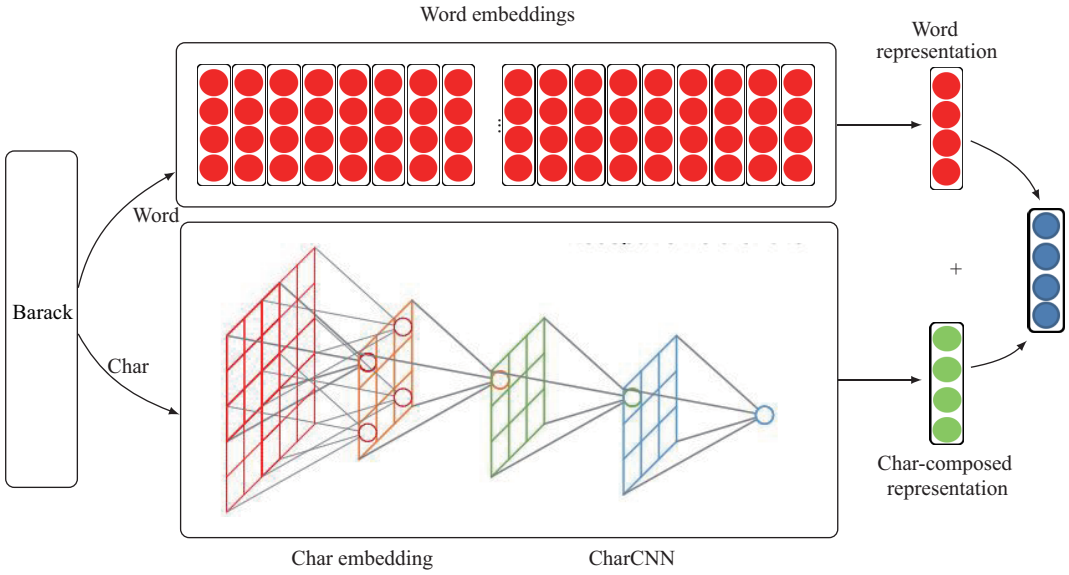

图 1 (网络版彩图) 词表示学习模型架构图

Figure 1 (Color online) The framework of character-based word representation learning

信息和上下文信息, 可以得到更好的词表示. 实体和关系的表示, 通过 BiLSTM [19] 模型对得到的词表示进行组合, 得到的隐层向量相加得到最终的表示.

对于一个给定的实体 e, 其表示 $\boldsymbol{e}$ 通过下面的模型得到

$$\boldsymbol{e} = \text{BiLSTM}(\text{Rep}(w_1^{\text{name}}), \ldots, \text{Rep}(w_n^{\text{name}})) + \text{BiLSTM}(\text{Rep}(w_1^{\text{type}}), \ldots, \text{Rep}(w_m^{\text{type}})), \tag{1}$$

其中 w_i^{name} 是 $\boldsymbol{e}$ 实体名的第 i 个单词; w_j^{type} 是 $\boldsymbol{e}$ 实体类型名的第 j 个单词; $\text{Rep}(w)$ 是单词的最终表示, 计算方法如下所示:

$$\text{Rep}(w) = \text{CNN}(\boldsymbol{c}_1^w, \ldots, \boldsymbol{c}_k^w) + \text{WE}(w), \tag{2}$$

其中 $\boldsymbol{c}_i^w$ 是单词 w 的第 i 个字符的向量表示; $\text{WE}(w)$ 是词 w 的词向量. 本文将实体的名字和类型表示相加作为实体的表示. 知识库中关系 r 的表示 $\boldsymbol{r}$ 则是通过对其描述文本使用组合模型得到, 计算方法如下所示:

$$\boldsymbol{r} = \text{BiLSTM}(\text{Rep}(w_1^{\text{relation}}), \ldots, \text{Rep}(w_m^{\text{relation}})), \tag{3}$$

其中 w_i^{relation} 是 r 对应的第 i 个单词.

本文基于已有的知识表示学习模型来将结构语义编码到文本的表示和组合模型中. 对于给定的三元组, 首先基于上述方法将实体和关系转换为向量表示, 然后训练知识库模型. 模型在训练的过程中, 不是更新某一个实体或关系的表示, 而是更新字符向量、词向量、CharCNN 模型和 BiLSTM 模型. 通过这种方式可以将知识库的结构信息作为一种约束加入到文本的表示学习中. 具体地, 针对给定三元组 (h, r, t), 首先根据文本表示模型得到对应的向量 $\boldsymbol{h}$, $\boldsymbol{r}$, $\boldsymbol{t}$. 然后使用打分函数来计算这个三元组的分值:

$$\text{score} = f(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}), \tag{4}$$

其中 f 是打分函数, 不同的模型具有不同的打分函数, 打分函数代表了该三元组的损失, score 值越小, 该三元组越可能为真. 本文使用目前应用最为广泛的 TransE [11] 模型作为知识表示的模型, 其打分函

数如下所示:

$$f(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\,)=|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}|_2^2. \tag{5}$$

知识库表示学习模型在训练时需要负例, 并让正例的分值大于负例. 但是知识库本身只包含正例, 本文借助于 TransR [29] 的方法, 对于给定的三元组, 随机替换其中的头部实体 h 或者尾部实体 t 来构建负例集合 $\mathbb{T}'$. 知识库表示模型的损失函数定义为

$$\text{loss}=[\text{margin}-(f(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})-f(\boldsymbol{h}',\boldsymbol{r},\boldsymbol{t}'))]_+, \tag{6}$$

其中 $(h,r,t)\in\mathbb{T}$ 是知识库正例集合; $(h',r,t')\in\mathbb{T}'$ 是构建的负例集合; margin 是控制参数. 直观上可以看作模型要求正例的置信度要比负例的置信度大至少 margin. 本文通过知识表示的学习, 将知识库的结构信息融合到文本的表示和组合模型中.

3.2 实体链接

本小节主要介绍在给定问句和候选实体的情况下, 如何确定该问句中所涉及的正确实体. 本文将问题形式化为给定一个包含 n 个词的问句 $Q=w_1,w_2,\ldots,w_n$, 本文首先通过图 1 所示的方式学习得到问句中每个词的表示. 基于得到的词表示, 使用 BiLSTM 组合模型对句子进行组合得到句子表示. 一个问句中往往包含多个词, 并且每个词对于实体的重要程度是不同的. 因此在学习句子的表示时, 本文引入了注意力机制, 该模型将每个候选实体的表示作为 attention 来监督句子的表示学习. 问句表示的计算方式如下所示:

$$\boldsymbol{Q}=\text{Comp}(w_1,w_2,\ldots,w_n)=\boldsymbol{h}_1\cdot\text{att}_{k1}+\boldsymbol{h}_2\cdot\text{att}_{k2}+\cdots+\boldsymbol{h}_n\cdot\text{att}_{kn}, \tag{7}$$

其中 w_i 为问句中第 i 个词; $\boldsymbol{h}_i$ 是句子第 i 个词对应的隐藏层的表示; att_{ki} 是第 i 个词对应的第 k 个实体的注意力权重, 基于下式计算得到.

$$\text{att}_{ki}=\frac{\exp(\text{score}(\boldsymbol{h}_i,\boldsymbol{e}_k))}{\sum_{i'}\exp(\text{score}(\boldsymbol{h}_{i'},\boldsymbol{e}_k))}, \tag{8}$$

其中 $\text{score}(\boldsymbol{a},\boldsymbol{b}\,)$ 是两个向量的相似度计算方法; $\boldsymbol{e}_k$ 是第 k 个实体对应的向量表示. 本文使用余弦相似度计算向量之间的相似度.

根据上述方法, 可以对每个问句以及对应的每个候选实体得到分布式表示. 问句与候选实体之间的相关度通过对应向量之间的余弦相似度来衡量. 这样相似度最大的候选实体即作为识别出来的实体.

3.3 关系识别

本小节主要介绍在给定问句和实体的情况下, 如何确定该问句中所涉及的关系. 在确定实体之后, 所有与该实体在三元组中共现的关系都作为候选关系. 候选关系的表示直接使用组合模型对其描述文本进行组合表示学习得到. 与实体类似, 相同的问句在涉及不同关系时每个词的重要程度是不同的. 例如, 给定问句 “what country was the film the debt from”, 对应的关系是 “film/film/country”, 则问句中的 “country” 和 “film” 对于识别该关系的重要性要大于其他词. 因此在关系识别中, 句子的表示学习也是用关系表示来作为注意力机制的依据. 其计算方法与实体链接类似, 如下:

$$\boldsymbol{Q}=\text{Comp}(w_1,w_2,\ldots,w_m)=\boldsymbol{h}_1\cdot\text{att}_{l1}+\boldsymbol{h}_2\cdot\text{att}_{r2}+\cdots+\boldsymbol{h}_m\cdot\text{att}_{lm}, \tag{9}$$

其中 Comp 模型与实体识别中的模型相同, 且参数共享; m 是该关系对应的词的个数; att_{li} 是问句中第 i 个词对应第 l 个关系的权重, 计算方法如下所示:

$$\mathrm{att}_{li} = \frac{\exp(\mathrm{score}(\boldsymbol{h}_i, \boldsymbol{r}_l))}{\sum_{i'} \exp(\mathrm{score}(\boldsymbol{h}_{i'}, \boldsymbol{r}_i))}. \tag{10}$$

3.4 模型训练

给定问句以及对应的三元组, 模型的目标是在给定问句的情况下, 找到该问句所涉及的实体及关系. 本文首先将知识库以及问句所涉及的所有文本构建词表, 然后初始化字符向量和词向量. 字符向量、词向量和字符级别的 CNN 组合模型在知识库表示和问答系统是共享的. 联合训练的总体步骤如算法 1 所示. 首先, 基于知识库中的三元组集合 $(h, r, t) \in \mathbb{T}$ 将三元组的结构语义信息编码到词向量、字符向量、CNN 字符组合模型和 BiLSTM 组合模型中. 其次, 使用得到的词向量、字符向量、CharCNN 和 BiLSTM 来训练问句的实体链接和关系识别模型. 最后模型迭代地执行知识表示学习和实体链接/关系识别.

算法 1 融合知识表示的知识库问答系统

Require: 三元组 $\mathbb{T}$, 问句 – 实体/关系对, 实体名、实体类型名、关系描述集合;
Ensure: 字符向量 $\mathbb{C}$, 词向量 $\mathbb{W}$, CNN 模型, BiLSTM 模型;
  根据三元组 $\mathbb{T}$ 构建负例集合三元组 $\mathbb{T}'$;
  构建词表、初始化向量及 CNN 和 BiLSTM 组合模型;
  while 未收敛或未到停止条件 **do**
    利用三元组、文本描述等信息, 训练知识表示模型;
    更新字符、词向量, 更新 CNN, BiLSTM 组合模型;
    利用问答数据、词向量和组合模型等训练基于知识库的问答系统;
    更新字符、词向量, 更新 CNN, BiLSTM 组合模型;
  end while

4 实验

4.1 实验数据

本文使用 SimpleQuestion [7] 数据集来验证方法的有效性. 该数据集包含超过 10 万个简单问题, 每个问题与 Freebase 知识库中的一个三元组对应, 可以用来回答该问题. 比如: 问句 “who produced change partners”, 对应三元组 (m.0s4hxt, music/recording/producer, m.0d7n9). SimpleQuestion 数据集划分为训练集、测试集和开发集, 分别包含 79910, 21678 和 10845 条数据. 本文使用 Freebase 的子集 FB2M 作为训练数据来进行知识表示的训练, 包含个约 2000000 实体和 6701 个关系. 为了加快训练速度, 本文对 FB2M 数据集进行预处理, 过滤掉所有不包含 SimpleQuestion 中实体或关系的三元组. 最终产生的知识库数据集包含 131062 个实体和 3059 个关系. 在知识库的表示学习中, 本文采用 Lin 等 [29] 提出的方法构建负例. 在进行实体替换时, 替换的实体应在相应的位置出现过, 且不能是已经存在知识库中的三元组. 例如, 当替换头部实体时, 该实体在训练集中作为头部实体出现过.

4.2 模型设置

本文基于 CNN 来实现对字符组合形成单词的表示, 使用 BiLSTM 实现对单词的组合, 可以得到

表 1 SimpleQuestion 数据集上的结果

Table 1 The results on SimpleQuestion dataset

Method	Precision (%)	Improvement
Bordes et al. [7]	62.7	%+15.3%
Yin et al. [8]	68.3	%+5.85%
Dai et al. [38]	62.6	%+15.5%
Golub and He [6]	70.9	%+1.97%
Lukovnikov et al. [37]	71.2	%+1.54%
Ours	**72.3**	–

实体、关系和问句的表示. 模型基于 Pytorch 实现, 采用 Adam 算法进行参数更新. 词向量和字符向量维度设置为 50; CNN 模型的词窗口设置为 3, 包括两个卷基层和两个 Max-Pooling 层和一个全链接层; BiLSTM 是双向 LSTM 模型, 输出维度设置为 100; 学习率初始化为 $\{0.1, 0.001, 0.0001\}$; TransE 模型中的 margin 设置为 $\{0.5, 1.0, 2.0\}$, batch size 设置为 $\{100, 500, 2000\}$, dropout 设置为 0.5, 在联合训练中, TransE 每次迭代 200 次, 知识库问答系统迭代 5 次, 总的迭代次数设置为 100 次. 模型利用 SimpleQuestion 的验证集进行提前停止 (early stop).

4.3 实验结果

本文与近期提出的最好的方法进行对比, 这些方法均基于 SimpleQuestion 数据集进行验证. 本文对比的方法均为端到端类型的方法, 包括 memory network [7], Yin 等 [8] 提出的基于 attention CNN 的方法, Golub 和 He [6] 提出的基于 character-attention 的 encoder-decoder 方法, Dai 等 [38] 提出的基于 RNN 的方法, 以及 Lukovnikov 等 [37] 提出的 word/character 方法. 对于利用多种组合和资源构建起来的方法, 因为其人工参与的程度或使用更多的资源等原因, 与本文不具有可比性 (例如 Dai 等 [38] 使用的 active linking 和 Yin 等 [8] 的 focused pruning). 评价指标采用准确率, 只有当识别出来的实体和关系都正确的情况下才判定为正确. 实验结果如表 1 所示. 从表 1 可以得到以下结论:

- 本文提出方法学习到统一底层表示, 文本和知识的表示都基于该表示得到, 解决了数据异构的问题;
- 与文献 [10] 的结果对比, 我们的方法取得了更好的结果, 基于统一的表示学习到的文本和知识表示之间的相似度能反映数据之间的语义相关性;
- 本文提出方法相对于其他端到端的模型取得了更好的结果.

如表 1 所示, 本文提出的方法在端到端类型的方法中取得了最好的效果 (最优结果用粗体表示). 可以看出本文提出的方法能够有效地利用知识表示来增强基于知识的问答系统任务.

4.4 详细分析

4.4.1 不同模块的作用

为了更好地判断知识库的字符表示、注意力机制、知识表示, 以及联合训练对知识问答系统的作用, 本文在 SimpleQuestion 数据集上进行多组实验来验证各个模块的效果. 这部分实验对比的模型包括: 采用随机的方法来选择实体和关系, 基于词向量 + BiLSTM 的模型, 在词向量 + BiLSTM 模型基础上增加 attention 机制, 在词向量 + BiLSTM + attention 的模型的基础上增加结构语义信息. 实验结果如表 2 所示.

表 2 不同模块的组合在 SimpleQuestion 数据集上的结果

Table 2 The results achieved based on different modules of our model

Number	Method	Precision (%)
1	Random (50)	2.1
2	WE + BiLSTM	43.1
3	WE + BiLSTM + CharCNN	62.3
4	WE + BiLSTM + CharCNN + attention	66.6
5	WE + BiLSTM + CharCNN + attention + KB Structure	71.3
5	WE + BiLSTM + CharCNN + attention + KB Structure + Jointly	**72.3**

表 3 未召回实体对准确率的影响

Table 3 The impact of lost entities for accuracy

Entity recall	Overall precision (%)	The precision after filtering out the non-recall entities	Improvement
88.1	72.3	82.1	%+12.3%

从表 2 可以得出如下结果: (1) CharCNN 可以得到比较大幅度的特征, 我们推测是因为 CharCNN 能够更好地处理生僻词、集外词, 学习到其有效的向量表示 (有 24121 个实体只在测试集中出现过, 没有在训练集中出现). (2) 通过 attention 机制, 模型可以找到句子中的关键词, 能够更好地定位到最相关的词汇, 从而帮助模型识别出正确的实体和关系. (3) 知识库的结构信息对基于知识的问答系统有比较明显的作用, 提升了 7% 左右的准确率. (4) 知识库表示和知识问答系统的联合训练能够提升模型的效果, 通过迭代训练可以进一步提升实体链接和关系识别的准确率. 上述实验验证了知识表示的作用, 本文提出的模型能够有效地将结构语义信息编码到文本表示中, 进而提升实体和关系识别的准确率.

4.4.2 错误分析

为了更好分析本文提出模型的适用范围以及局限性, 本小节对出现的主要错误类型进行了分析. 发现通过 n-gram 做实体匹配之后, 有时候并不能得到正确的候选实体, 这些问句所涉及的实体不能召回. 而当正确的实体没有召回时, 模型显然不能得到正确的结果. 因此本文提出的模型对于候选实体选择的方法有较强的依赖性, 其召回的结果直接影响模型的效果. 表 3 给出了模型在不同条件下的准确率情况.

表 3 显示, 当过滤掉未召回的实体之后, 模型的准确率提升了 12.3%. 因此, 我们的模型可以通过更好的候选实体生成技术来进一步提升准确率.

5 结束语

本文通过将知识表示融合到文本的表示和组合中的方法, 有效提高了基于知识库的问答系统的效果. 实验表明知识表示能够有效地提升问句中实体链接和关系识别的效果, 字符级别的组合、注意力机制和联合训练等方式都具有提高识别问句实体和关系准确率的作用. 本文提出的模型在基于知识库的简单问题数据集上取得了较好的结果. 通过详细分析系统预测的错误, 发现该模型比较依赖于候选实体的生成, 简单的基于 n-gram 匹配的候选实体生成模型的召回率有限, 因此模型的准确率有望进一步提升. 另外, 由于语言的多样性, 组合模型有时不能很好地处理相同关系的不同说法, 导致不能正

确地识别关系. 下一步工作, 我们希望使用同义词扩展等功能提高候选实体生成的召回率. 同时, 我们还计划使用复述数据和语义解析技术来增强模型对于语言多样性的处理能力.

参考文献

1 Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of ACM SIGMOD International Conference on Management of Data, 2000. 1247–1250

2 Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge. In: Proceedings of International Conference on World Wide Web, 2007. 697–706

3 Miller G A. WordNet: a lexical database for English. Commun ACM, 1995, 38: 39–41

4 Zettlemoyer L S, Collins M. Learning context-dependent mappings from sentences to logical form. In: Proceedings of the Meeting of the Association for Computational Linguistics, 2009. 976–984

5 Bos J, Clark S, Steedman M, et al. Scaling semantic parsers with on-the-fly ontology matching. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013. 1545–1556

6 Golub D, He X D. Character-level question answering with attention. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016. 1598–1607

7 Bordes A, Usunier N, Chopra S, et al. Large-scale simple question answering with memory networks. 2015. ArXiv:1506.02075

8 Yin W P, Yu M, Xiang B, et al. Simple question answering by attentive convolutional neural network. 2016. ArXiv:1606.03391

9 Zhou B T, Sun C J, Lin L, et al. LSTM based question answering for large scale knowledge base. Acta Sci Natl Univ Pekinensis, 2018, 54: 286–292 [周博通, 孙承杰, 林磊, 等. 基于 LSTM 的大规模知识库自动问答. 北京大学学报: 自然科学版, 2018, 54: 286–292]

10 Hao Y C, Zhang Y Z, Liu K, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In: Proceedings of Meeting of the Association for Computational Linguistics, 2017. 221–231

11 Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 2787–2795

12 Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, 2008. 160–167

13 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 3111–3119

14 Pennington J, Socher R, Manning C D. Glove: global vectors for word representation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2014. 1532–1543

15 Zhang X, Zhao J B, Lecun Y. Character-level convolutional networks for text classification. In: Proceedings of International Conference on Neural Information Processing Systems, 2015. 649–657

16 Ling W, Luís T, Marujo L, et al. Finding function in form: compositional character models for open vocabulary word representation. 2015. ArXiv:1508.02096

17 Mitchell J, Lapata M. Composition in distributional models of semantics. Cogn Sci, 2010, 34: 1388–1429

18 Paulus R, Socher R, Manning C D. Global belief recursive neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2888–2896

19 Melamud O, Goldberger J, Dagan I. Context2vec: learning generic context embedding with bidirectional LSTM. In: Proceedings of SIGNLL Conference on Computational Natural Language Learning, 2016. 51–61

20 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017. ArXiv:1706.03762

21 Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data. Mach Learn, 2014, 94: 233–259

22 Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases. In: Proceedings of AAAI Conference on Artificial Intelligence, 2011

23 Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion. In: Proceedings of International Conference on Intelligent Control and Information Processing, 2013. 464–469

24 Sutskever I. Modelling relational data using bayesian clustered tensor factorization. In: Proceedings of Advances in Neural Information Processing Systems, 2009. 1821–1828

25 Yang B, Yih W T, He X. Learning multi-relational semantics using neural-embedding models. 2014. ArXiv:1411.4072

26 Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction. In: Proceedings of International Conference on Machine Learning, 2016. 2071–2080

27 Wang Z, Zhang J W, Feng J L, et al. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014. 1112–1119

28 Fan M, Zhou Q, Chang E, et al. Transition-based knowledge graph embedding with relational mapping properties. In: Proceedings of Pacific Asia Conference on Language, Information and Computation, 2014. 328–337

29 Lin Y K, Liu Z Y, Sun M S, et al. Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of Conference on Artificial Intelligence, 2015. 2181–2187

30 Lin Y K, Liu Z Y, Luan H B, et al. Modeling relation paths for representation learning of knowledge bases. 2015. ArXiv:1506.00379

31 Wang Q, Wang B, Guo L. Knowledge base completion using embeddings and rules. In: Proceedings of International Conference on Artificial Intelligence, 2015. 1859–1865

32 Rocktäschel T, Singh S, Riedel S. Injecting logical background knowledge into embeddings for relation extraction. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics, 2015. 1119–1129

33 Guo S, Wang Q, Wang L H, et al. Jointly embedding knowledge graphs and logical rules. In: Proceedings of Conference on Empirical Methods on Natural Language Processing, 2016. 192–202

34 Milajevs D, Kartsaklis D, Sadrzadeh M, et al. Evaluating neural word representations in tensor-based compositional settings. 2014. ArXiv:1408.6179

35 Milajevs D, Kartsaklis D, Sadrzadeh M, et al. Evaluating neural word representations in tensor-based compositional settings. 2014. ArXiv:1408.6179

36 Zettlemoyer L S, Collins M. Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. 2012. ArXiv:1207.1420

37 Lukovnikov D, Fischer A, Lehmann J. Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of International Conference on World Wide Web, 2017. 1211–1220

38 Dai Z H, Li L, Xu W. CFO: conditional focused neural question answering with large-scale knowledge bases. 2016. ArXiv:1606.01994

Knowledge-representation-enhanced question-answering system

Bo AN[1,2*], Xianpei HAN[2] & Le SUN[2]

1. *University of Chinese Academy of Sciences, Beijing* 100190*, China*;
2. *Institute of Software, Chinese Academy of Sciences, Beijing* 100049*, China*
* Corresponding author. E-mail: anbo@iscas.ac.cn

Abstract A knowledge-based, question-answering system can automatically answer natural language questions using triples in a knowledge graph. Simple questions are the most common type of questions which can be answered by a single triple. However, answering simple questions is still challenging when faced with a large-scale knowledge graph. Currently, most end-to-end models learn the distributional representations of entities, relations, and questions, and compute semantic relevance based on these representations. These models ignore the structural information of knowledge graphs, which is important for entity linking and relation recognition in questions. In this paper, we propose a unified representation learning method for text and knowledge to learn the representations of questions, entities, and relations. The structural information of a knowledge graph constrains the word representation and the word composition model. Experimental results over a knowledge-based question-answering system show that the question representation and the knowledge representation learned based on the unified representation can achieve a better performance.

Keywords question answering system, knowledge base, word composition, knowledge representation, text representation

Bo AN was born in 1986. He received his B.S degree in computer science and technology from Dalian University of Technology, Dalian, China, in 2011. Currently, he is a Ph.D. candidate at the University of Chinese Academy of Sciences. His research interests include knowledge representation and natural language understanding.

Xianpei HAN received his Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Currently, he is a professor at the Institute of Software, Chinese Academy of Sciences. His research interests center on large-scale knowledge base population using heterogeneous information sources.

Le SUN received his Ph.D. degree from Nanjing University of Science & Technology. From 1998 to 2000, a post-doc researcher at Chinese Information Processing Center, Institute of Software, Chinese Academy of Sciences. Currently, he is a professor at the Institute of Software, Chinese Academy of Sciences. His major research interests include knowledge-based natural language understanding and Chinese information processing.